DSC 510: Healthcare Data Science

Dataset:

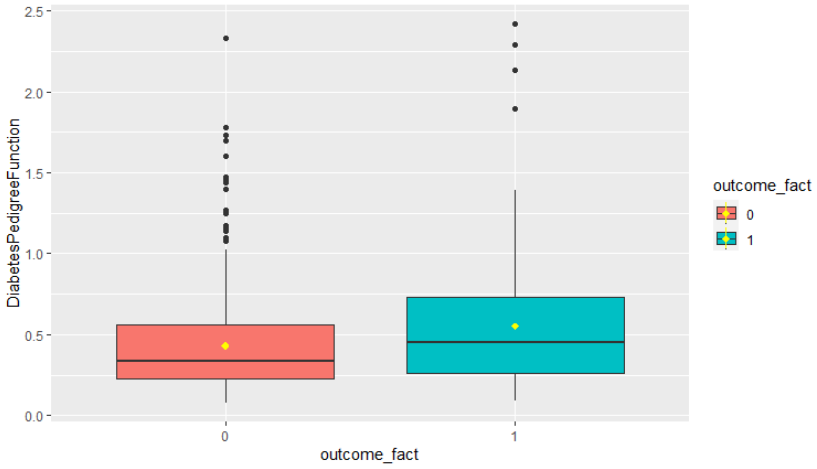
Pima Indian dataset from Kaggle has 768 observations and 9 variables. The data consists of one binary target variable, diabetes outcome which indicated with a 0 (non-diabetic) or 1 (diabetic). There are 8 different medical predictor variables: number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. While the data had no null values, some cases had 0 in certain columns that would be improbable, if not impossible. These likely indicated that the test was not conducted for a BMI of 0 is not possible. Therefore, the list wise deletion was used to clean the dataset. In the end, the data set contained complete 392 observations and 9 variables, 8 independent and 1 dependent.

Research question: Is there a difference between diabetes test results for the pedigree function?
H0: There is no difference between diabetes test results for the pedigree function.
Ha: There is a difference between diabetes test results for the pedigree function.

I think the answer for exploratory analysis will be there is a difference between diabetes test results for pedigree function since the pedigree function is the likelihood score of diabetes based on family history and the cause of diabetes can be genetic and other factors like diet. Therefore, people who have diabetes will likely have a higher score in pedigree function.

The Shapiro-Wilk normality test is used to test normality and the Wilcox test is used to test if there is a difference between diabetes test results for the diabetes pedigree function. Based on the previous correction test, the diabetes pedigree function has weak positive correlations with Outcome. The Shapiro-Wilk normality test shows that the pedigree function is not a normal distribution, and the histogram of the pedigree function that is grouped by Outcome is not normally distributed, either. Therefore, utilizing the Wilcox test to test the hypothesis is more appropriate. From the Wilcox test results, the p-value is 0.000001197 that is less than 0.05. It indicates the null hypothesis is rejected, so there is a difference between diabetes test results for the diabetes pedigree function.

The box plot shows that there is a difference between diabetes test results for the diabetes pedigree function since they have different mean, median, interquartile range, and max. Patients who are diabetics are likely to have higher scores of pedigree function.



Research question :What explains whether a patient will have diabetes? (logistic regression)
After I drop the rows that have values equal to 0 in Glucose, BMI, BloodPressure, SkinThickness, and Insulin, the observations become 392 and the correlation test shows that Glucose and Outcome have a moderate correlation (0.52). The rest variables have weak positive correlations with Outcome. The result from logistic regression shows that Glucose, BMI, and pedigree function have positive influence on Outcome and they are also statistically significant to determine whether someone will have diabetes since their p-values are less than 0.05. The p-value of Glucose is <0.001, the p-value of BMI is 0.01, and the p-value of DiabetesPedigreeFunction is 0.008. The p-value of Age, Insulin, SkinThickness, BloodPressure, and Pregnancies are greater than 0.05 which mean they are not statistically significant to determine whether someone will have diabetes.

Characteristic	OR[1]	95% CI[1]	p-value
Pregnancies	1.09	0.97, 1.21	0.14
Glucose	1.04	1.03, 1.05	<0.001
BloodPressure	1.00	0.98, 1.02	>0.9
SkinThickness	1.01	0.98, 1.05	0.5
Insulin	1.00	1.00, 1.00	0.5
BMI	1.07	1.02, 1.13	0.010
DiabetesPedigreeFunction	3.13	1.38, 7.37	0.008
Age	1.03	1.00, 1.07	0.065

[1] OR = Odds Ratio, CI = Confidence Interval

Discuss the limitations of the analyses
One of the limitations of my analysis is how the pedigree function is calculated. The data doesn't have missing values but there are several variables like BMI, glucose, blood pressure, and insulin values that are equal to 0. It is difficult to know if the pedigree function is zero in this dataset really represents the patients do not have any family members who have diabetes. Since I chose listwise deletion, 40 percent of observations that have values equal to 0 in BMI,

glucose, blood pressure, and insulin got deleted. The data became small, so it'd affect the machine learning results.

Machine learning-XGBoost:

XGBoost (eXtreme Gradient Boosting) was utilized to predict if a patient gets diagnosed with diabetes or non-diabetes. XGBoost is a type of enabling decision tree models which can be implemented for classification. First, we split 80 % of the data into the training set and 20% into the test set. There are 77 observations in the test set, and 25 of them get diagnosed with diabetes. The result from the initial model when we set the learning rate to 1 (eta=1) had low accuracy. After applying grid search to tune the parameters to increase the accuracy, such as decreasing the learning rate, increasing max depth of a tree, and increasing the number of iterations, the sensitivity and accuracy got increased. The parameters of the best model for this method was setting 20 for the maximum depth of a tree, the learning rate was 0.01, and the maximum number of iterations was 100. The result achieved an accuracy of 88.31% with 95% Confidence Interval (0.7897, 0.9451) in predicting diabetes, the sensitivity was 84% , the specificity was 90.38%, the positive predictive value was 80.77%, and the negative predictive value was 92.16% . The importance of variables of this model showed that Glucose (38%) had the most influence in predicting diabetes, followed by insulin, age, and BMI which were almost or above an importance of 10%.

Table : Confusion Matrix of XGBoost model

Confusion Matrix	**Actual Class**	
Prediction	Non-diabetic	diabetic
Non-diabetic	47	4
diabetic	5	21

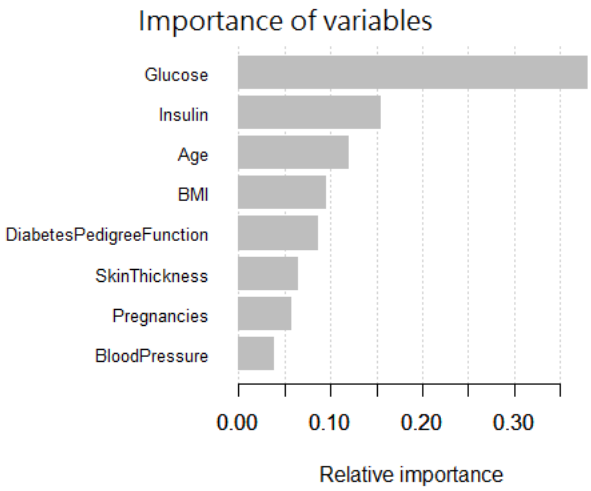


Figure : Relative Importance of features with respect to diabetes outcome in Pima Indian dataset

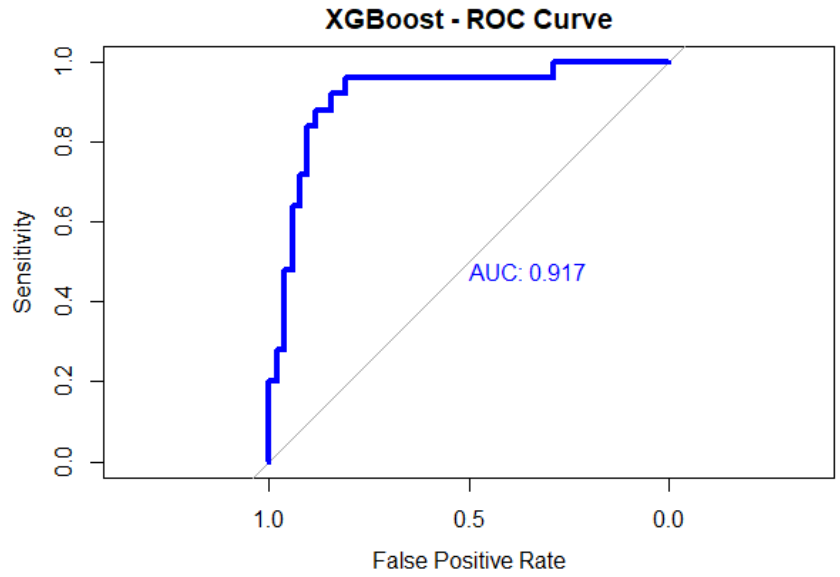


Figure : AUC and ROC Curve of Diabetes of XGBoost classifier from the test set

XGBoost model showed that Glucose is the most critical factor in predicting diabetes, which corresponds to glucose level is the main criteria to determine if a patient has diabetes. The applications of these two models can be used in monitoring glucose level and BMI of the Pima Indian women who might be at high risk of being diagnosed with diabetes and help them by increasing nutritional food sources and encouraging them to increase levels of physical activity, so their risk of getting diabetic can be lower.